N
523

HW 25/38

PAPER ON STATISTICS OF REPETITIONS

by A.M. Turing

Statistics of Repetitions

In order to be able to obtain reliable estimates of the value of given repeats we need to have information about repetition in plain language. Suppose for example that we haveplaced two messages together and that we find repetitions consisting of ~~two~~ a tetragramme, two bigrammes and fifteen singel letters, and that the total 'overlap' was 105, i.e. that the maximum possible number of repetitions which could be obtained by altering letters of the messages is 105: suppose also that the lengths of the messages are 200and 250: in such a case what is the probability of the fit being right, no other information about the days traffic being taken into consideration, but information about the character of the unenciphered text being available in considerable quantity?

In theory this can be solved as follows. We take a vast number of ~~messages~~ typical decodes, say 10^{10^6}, and from these we select all of length 200 and all of length 250. We encipher all of these messages at all possible positions on the machine (neglecting for simplicity the complication dueto different daily keys). We then compare each message 200 long with each 250 long in such a way as to get an overlap of 105 as with the fit under consideration. From the resulting comparisons we pick out just those cases where the repetitions have precisely the same form as in the case in question. This set of comparisons will be called the 'relevant' comparisons. Among the relevant comparisons there will be some which are 'right' comparisons, i.e. where corresponding letters of the two messages were enciphered with the same position of the machine. The probability that our original fit was right can now be expressed in the form

$$\frac{\text{Number of right relevant comparisons}}{\text{Total number of relevant comparisons}}$$

The work involved in this theoretical method can be vastly reduced if we make a few harmless assumptions. In the first place if we assume that the encipherment keys at the various positions of the machine are 'hatted' we can calculate the number of relevent wron g comparisons. Suppose the total number of repeated letters in th e ase in question is R, then

$$\frac{\text{Number of relevent wrong comparisons}}{\text{Total number of wrong comparisons}} = \left(\frac{1}{26}\right)^{R} \left(\frac{25}{26}\right)^{L-R}$$

For th calcul tion of the number of relevent right comparisons we have to make other assumptions. The sort of assumption that we need is that a repetition in one place is not made any the more or less likely by a repetition elsewhere. Actually this assumption would not be quite true, as it clearly does not hold in the case of a jacent letters. For most practical purposes I think the following assumption is sufficiently near to the truth:- If we know that at a certain point P there is not a repetition, th en knowledge that there is or is not a repetition at a point A before P does not make a repetition at a point B after P either more likely or less likely. ~~Withxthisxassumptionxwex xshouldxgetxthexrightxdistributionxofxrepetitionxfiguresxfor xcomparisons~~ The probability of a repetition at any point is also independent of its distance from the ends of either message. With these assumptions we could get the right distribution of numbers of comparisons between ther various repetition figures if we assume the repetition figures for the comparisons constructed in this way. We are given an urn containing a large number of ~~xxxxx xfxpaper~~ cards, some bearing the words 'no repeat', some 'simple repeat', some 'bigramme', some 'trigramme', and so on. To construct a random sample of repetition figures for comparisons of given length we make a series of draws from the urn.

The first few draws determine the repetition figure for the first comparison, the next few for the next comparison, and so on. When we draw 'no repeat' we have to add a 'o' to the repetition figure, when we draw 'simple repeat'we add 'xo', for 'bigramme' we add 'xxo' and so on. When we have got to the right length of overlap required the comparison is completed and our next draws refer to the next comparison, If it happens that the right length is never reached because we'jump past it' then we scrap that comparison, and go on to the next. As an example suppose that we are making comparisons with an overlap of 12, and that our first draws are 'tetragramme', 'no rep', 'no rep', 'no rep', 'bigramme', 'no rep', 'trigramme', 13- gramme, then 'no rep' 13 times, our first two comparisons will have the repetition figures

xxxxooooxxoo

oooooooooooo

the one starting xxxo being rejected because we never reach the right lenght of overlap. (This arrangement requires that every repetition figure should end with o, and therefore the genuine repetition figure should be obtained by crossing this off: but I shall not be too meticulous about details arising from the ends of the comparisons). The number of draws required to produce a given figure is the number of non repeating letters, i.e. the overlap less the number of repeating letters. With our convention about crossing off the last letter we have to add 1.

Two problems arise from this picture

1) How do we calculate the correct proportions of cards in the urn?

2) Given the proportions of the cards in the urn, how do we calculate the number of right relevent comparisons, and hence the probability of a given fit?

The correct $\frac{1}{p}$ proportion of the cards in the urn can be calculated from the actual distribution of repetitions in the case of messages correctly set, or, what comes to the same thing, in messages unenciphered and arbitrarily set. Let us suppose that we have a large number of such comparisons of unenciphered messages, and that the messages are sufficiently long that complications arising from the ends of the messages can be neglected. The proportion of cards bearing the words 'simple repeat','bigramme', 'trigramme' etc must obviously be in the same ratio as the number of corresponding repeats in our comparisons. The number of 'no repeat' cards will be calculated slightly differently as we have to subtract one case of 'no repeat' for each sequence of repeating letters.

To get the best value from given material we naturally make every possible comparison. If we do this the right number of repetitions can be calculated quite easily without actually making the comparisons. Theoretically we can imagine the complete set of comparisons made in this way. First of all we write out all the decodes (say 50 of them) one after another round a circle: suppose that the number of letters on this circle is N. The whole is then repeated on a concentric circle. All possible comparisons can be made by rotating the one circle with respect to the other. From these we have to remove the comparison in which the circles are not rotated at all, for obvious reasons. Also when the rotation is more than 180^o we get essentially the same comparison as one with less than 180^o. The net effect of this, taking into account also the special case of exact 180^o rotation, is that the total overlap of all the comparisons is $\frac{N(N-1)}{2}$. Now let us consider for

example the total number of tetragramme repeats in all these comparisons. These can be divided into the repeats arising from AAAA those from AAAB .... those from ZZZZ, the largest contribution arising presumably from such tetragrammes as EINS. The number of tetragrammes arising from EINS consists of the number of <u>pairs of hexagrammes</u> such as QEINSR, VEINSW in which the first letters of each are different, the last different, and the remainder spell EINS. This number of pairs we will call the '<u>actual</u> number of tetragramme repeats arising from EINS'. The 'actual number of tetragramme repeats' is obtained by summing over AAAA,AAAB,...,EINS,...ZZZZ. This 'actual' number is not easily calculated directly, but we can more easily obtain the 'apparent number of tetragramme repeats', and this leads to the actual number. The '<u>apparent</u> number of tetragramme repeats arising from EINS' is defined to be the number of pairs of occurrences of EINS in the material, and 'the apparent number of tetragramme repeats' is defined by summation. We can also define the apparent number of tetragramme repeats in a comparison as the number of different series xxxx in the comparison. Thus a heptagramme repeat gives four apparent tetragramme repeats. The actual number of repeats can be calculated from the apparent in this way. Let M_r be the apparent number of r-grammes, and N_r the actual number. Then

$$M_r = N_r + 2N_{r+1} + 3N_{r+2} + \dots$$

So that

$$M_r - M_{r+1} = N_r + N_{r+1} + N_{r+2} + \dots$$

$$N_r = (M_r - M_{r+1}) - (M_{r+1} - M_{r+2}) = M_r - 2M_{r+1} + M_{r+2}$$

It is therefore sufficient to calculate only apparent numbers and to carry these two stages further than we want to go with the actual numbers. In practice octagramme repeats are so certain to be right that it will be sufficient tohave statistics only as far as heptagrammes. We therefore need statistics of apparent numbers of repeats as far as 9-grammes. To get these numbers of apparent repeatsit is sufficient to take all the 9-grammes in the material (i.e. on the circle) an d to put th em into alphabetical order. This can be done very cojveniently by say Hollerith. The number of trigramme repeats can then be found very simply (although with a good deal of labour) by considering only the first three letters of each 9-gramme. Suppose we denote by t a typical trigramme and by n_t the n umber of its occurrences, then the apparent number of trigramme repeatsis $\sum_t \frac{n_t(n_t-1)}{2}$.

~~In our later calculations it is convenient also to regard the comparisons in the wrong places as also constructed by drawing from anoth er urn. The proportions in this urn can in theory be calculated in th e same way, but from messages consistin g of hatted series of letters. In this case the apparent proportion of r-grammes is~~ $\cancel{\frac{1}{26^r}}$ ~~so that the actual proportion of r-grammes is~~ $\cancel{\frac{25^2}{26^{r+2}}}$.

When calculating the proportions of cards in th e urn we must remember that the total number of cards is not $\frac{N(N-1)}{2}$ but is less than this by $\sum r N_r$.

In our later calculations it is convenient to regard the comparisons in wrong places as also constructed by drawing from an urn. In this case we easily discover see that the apparent number of r-grammes is $\frac{N(N-1)}{2}\frac{1}{26^r}$, and from this we deduce that th e actual proportion of r-gramme cards is $\frac{25}{26^{r+1}}$ and of no repeat cards is $\frac{25}{26}$.

We now turn to the problem of calculating the probability of a given fit when we know the proportion α_r of r-gramme cards in the urn for each r. The calculation is going to be slightly complicated by the convention which we introduced, that not all drawings can lead to a comparison. We have therefore to calculate the proportion of draws which do lead to a comparison, i.e. in which the length does not overshoot the mark. The answer is that as the length of overlap tends to infinity the proportion tends to $\dfrac{1}{1+\sum r\alpha_r}$; in the case of hatted material this is $\frac{25}{26}$.

Now put $A = 1-\sum\alpha_r$. Consider a repetition figure in which there are k_r r-grammes. ~~Put~~ Let the overlap be L. The number of ~~draws which give no repeat is~~ 'no repeat' cards drawn is $L+1-\sum(r+1)k_r$. The proportion of right draws which are felevent is

$$A^{L+1-\sum(r+1)k_r}\prod_{r=1}^{\infty}\alpha_r^{k_r}$$

and the proportion of the right comparisons which are relevant is (assuming L reasonably large)

$$(1+\sum r\alpha_r)\,A^{L+1-\sum(r+1)k_r}\prod_{r=1}^{\infty}\alpha_r^{k_r}$$

Similarly calculating with the urn whose proportions were made up from hatted material we find for the proportion of wrong ~~draws~~ comparisons which are relevent

$$\frac{26}{25}\left(\frac{25}{26}\right)^{L+1-\sum(r+1)k_r}\prod_{r=1}^{\infty}\left(\frac{25}{26^{r+1}}\right)^{k_r}$$

Hence the odds* on our fit are

$$q = \lambda\,\frac{25(1+\sum r\alpha_r)}{26}\left(\frac{26A}{25}\right)^{L+1-\sum(r+1)k_r}\prod_{r=1}^{\infty}\left(\frac{26^{r+1}\alpha_r}{25}\right)^{k_r}$$

where λ is the a priori odds. This is most conveniently written as

$$\log q = \log\lambda + \sum\mu_r k_r - \nu L + \log(1-\sum\alpha_r)(1+\sum r\alpha_r)$$

where $\mu_r = \log\dfrac{\alpha_r 26^{r+1}}{25} - (r+1)\log\dfrac{26A}{25}$ and $\nu = \log\dfrac{25}{26A} \doteqdot \sum\alpha_r - 2/81$

*The odds on an event are defined to be the probability of the event divided by the probability of its negation

In the case of overlap zero there is a discrepancy of $\log(1-\Sigma\alpha_r)(1+\Sigma r\alpha_r)$ due to th e overlap not being long. This term is in any case microscopic.